Eur J Clin Microbiol Infect Dis (2014) 33:983–988
DOI 10.1007/s10096-013-2036-0

ARTICLE

Implementation of Fourier transform infrared spectroscopy for the rapid typing of uropathogenic *Escherichia coli*

S. E. Dawson · T. Gibreel · N. Nicolaou · H. AlRabiah · Y. Xu · R. Goodacre · M. Upton

Received: 28 November 2013 / Accepted: 10 December 2013 / Published online: 8 January 2014

Abstract In this paper, we demonstrate that Fourier transform infrared (FT-IR) spectroscopy is able to discriminate rapidly between uropathogenic *Escherichia coli* (UPEC) of key lineages with only relatively simple sample preparation. A total of 95 bacteria from six different epidemiologically important multilocus sequence types (ST10, ST69, ST95, ST73, ST127 and ST131) were used in this project and principal component-discriminant function analysis (PC-DFA) of these samples produced clear separate clustering of isolates, based on the ST. Analysis of data using partial least squares-discriminant analysis (PLS-DA), incorporating cross-validation, indicated a high prediction accuracy of 91.19 % for ST131. These results suggest that FT-IR spectroscopy could be a useful method for the rapid identification of members of important UPEC STs.

Introduction

Urinary tract infection (UTI) is the most common laboratory-confirmed bacterial infection encountered in medical practice in Europe and North America. Uropathogenic strains of *Escherichia coli* (UPEC) are the predominant cause of UTI and account for 70–95 % of community-acquired cases and approximately 50 % of hospital-acquired cases [1, 2]. UPEC differs from the majority of other enteric *E. coli* in terms of serotype, virulence-associated traits, phylogenetic grouping and degrees of multidrug resistance [3]. Recently, the management of UTIs has become increasingly challenging as a result of emerging resistance to most first-line antimicrobial agents [4]. Contributing to this increase are certain successful UPEC clones, including ST131, which is characterised as belonging to serotype O25b:H4, and the production of extended-spectrum β-lactamase and fluoroquinolone resistance [5, 6], as well as ST69, which is significantly associated with trimethoprim resistance [7].

Bacterial typing is a valuable tool in both outbreak investigation and epidemiological surveillance. A number of techniques have been deployed in order to try and reduce the time required for the identification of the causative organism in UTI. These include matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry [8], ultraviolet resonance Raman (UVRR) spectroscopy [9] and Fourier transform infrared (FT-IR) spectroscopy [10, 11].

FT-IR spectroscopy has become a useful technique for the discrimination of cultured bacteria. It has been successfully utilised for the differentiation of *Enterococcus* and *Streptococcus* species [12, 13], and is able to discriminate between different strains of *E. coli*, despite their high chemical similarity [12], and we have recently described method development for FT-IR allowing differentiation between quinolone-resistant and -sensitive members of ST131, and separation of these isolates from a small group of non-ST131 strains [11]. Here, we expand the study to demonstrate the prediction of UPEC of other key STs, using a large strain collection.

There are many advantages of using FT-IR for the identification of bacteria over traditional biochemical or molecular techniques, namely, the rapidity of the technique (~60 s to

S. E. Dawson · T. Gibreel · M. Upton
Microbiology and Virology Unit, School of Medicine, University of Manchester, Manchester, UK

N. Nicolaou · H. AlRabiah · Y. Xu · R. Goodacre
School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, Manchester, UK

M. Upton (✉)
School of Biomedical and Healthcare Sciences, Plymouth University, Portland Square, Drake Circus, Plymouth PL4 8AA, UK
e-mail: mathew.upton@plymouth.ac.uk

M. Upton
Formerly: School of Medicine, Stopford Building, University of Manchester, Manchester M13 9PL, UK

obtain spectra), automation, relatively low running costs and simple sample preparation [14].

Materials and methods

Bacterial specimens

A collection of 95 UPEC isolates belonging to six of the most common sequence types (STs) were collected between 2008 and 2010 from urine samples obtained from Manchester Royal Infirmary, Manchester, UK. UPEC strains had previously been typed using multilocus sequence typing (MLST) [15]. Nine isolates were selected from ST10, ST69, ST73, ST95 and ST127, and 50 isolates were of the ST131 lineage.

Sample preparation for FT-IR

Two full loops of biomass from a Columbia agar (Oxoid, Basingstoke, UK) plate were carefully collected and suspended in 500-μL aliquots of sterile physiological saline (0.89 % NaCl). Each sample was centrifuged at 13,000×*g* and washed three times with 500 μL sterile saline before being stored as a pellet at −80ºC. These aliquots were subsequently used for FT-IR analysis.

High-throughput screening (HTS) FT-IR spectroscopy

Samples were slowly defrosted on ice and the bacterial suspensions were adjusted to an optical density at 600 nm (OD_{600}) of between 0.4 and 1.2. Aliquots (30 μL) of each of the 95 bacterial samples were spotted onto a 96-well ZnSe plate (Bruker Optics Ltd., Coventry, UK) and oven dried at 40 °C for 30 min.

HTS FT-IR spectroscopic analysis was carried out using the microplate module of a high-throughput scanner (HTS-XT™) attached to an Equinox 55 spectrometer (Bruker Optics Ltd.) [14]. The spectrometer was fitted with a deuterated triglycine sulfate (DTGS) detector and controlled with Opus 4 software, via MS Windows on an IBM-compatible PC. As in previously described methods, FT-IR spectra were collected over the mid-IR wavelength range of 4,000–600 cm^{-1} and 64 scans were co-added and averaged [16, 17]. Three spectra (machine replicates) were collected from separate locations of each sample and the experiment was run in duplicate (two biological replicates), which resulted in a total of 570 spectra being collected. The FT-IR data were then converted and analysed using MATLAB 2010a (The MathWorks Inc., Natick, MA, USA).

Data analysis

Data were exported from the Opus software and imported into Matlab. The spectra were then normalised by using the standard normal variate (SNV) [18], a commonly used method to remove the light-scattering effect in the FT-IR spectra. In addition, the samples were analysed on two different FT-IR plates and the difference in the spectra between the two plates were clearly visible (data not shown). A piecewise direct standardisation (PDS) [19] model was built based on the common samples between the two plates and then this model was applied to the whole data set to remove the variation caused by using different plates.

Cluster analysis on the SNV and PDS pre-processed data was carried out in two stages using different algorithms. The unsupervised clustering method principal component analysis (PCA) was performed first to reduce the dimensionality of the multivariate data whilst preserving the variance [14]. PCA transforms the original set of variables to a new set of uncorrelated variables called principal components (PCs); these are ordered so that the first few retain the majority of the variation present in all of the original variables [20]. The PCs are then plotted against each other to visualise the natural clusters within the data. This was followed by discriminant function analysis (DFA), which is a supervised technique that discriminates data based on the retained PCs with a priori knowledge of which spectra are replicates. This is an unbiased method, as the provided priori knowledge only involved samples that were replicates. DFA maximises between-class variance and minimises within-class variance [21], and such a model with a priori knowledge of which samples were replicates minimises the variation caused by the experiment. Thus, the natural clustering trend due to the biological factors (e.g. phenotype of isolates of different STs) may better be presented, compared to the results of PCA.

In addition to PCA and DFA, partial least squares for discriminant analysis (PLS-DA) [22] was employed to build a classification model to predict the ST of the FT-IR spectra collected from the samples unknown to the model. This was done by using a bootstrapping resampling procedure [23]. In bootstrapping resampling, suppose there were n samples in total, n samples were randomly chosen with replacement (i.e. one sample could be chosen more than once). These samples were used as the training set and the remaining samples which had not been chosen were used as the test set (on average, 63.2 % of samples would be chosen as the training set and the remaining 36.8 % samples as the test set in each bootstrapping resampling). In our study, the bootstrapping resampling was based on biological replicates, i.e. all the samples of the same biological replicates were selected together, thus n=190 (95 isolates×2 biological replicates). The PLS-DA model was built on the training set and the number of PLS components was selected by using a k-fold cross-validation performed on the training set only, where k is the number of biological replicates in the training set. The model was then applied to the test set to predict their STs. This procedure was repeated 1,000 times and the prediction accuracies were averaged to show the generalisation performance of the PLS-DA model.

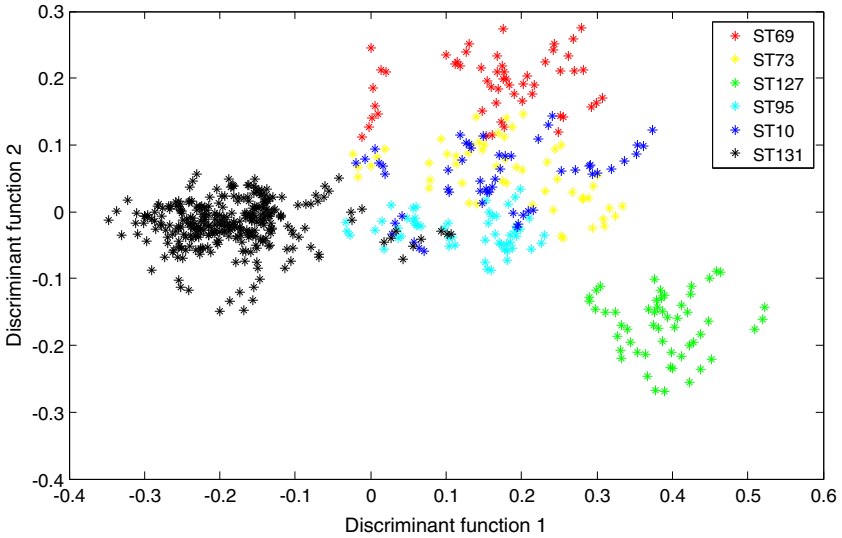

Fig. 1 Principal component-discriminant function analysis (PC-DFA) plot of high-throughput Fourier transform infrared (FT-IR) spectra for uropathogenic *Escherichia coli* (UPEC). PCs 1–20 (accounting for 99.12 % of the total variance) were used by the DFA algorithm with a priori knowledge of machine replicates (50 isolates of ST131, nine isolates from each of the other five STs examined) and, so, the fact that isolates cluster together is due to natural similarities, rather than ones artificially imposed by the DFA algorithm. Each different colour represents a different sequence type (ST), as shown by the figure legend. Clear clustering for ST127 and ST131 can be seen

Results

Initially, the raw spectra that were produced from FT-IR spectroscopy were visually examined (data not shown). These FT-IR spectra for the 95 UPEC isolates showed broad and complex contours, with some inevitable baseline shifts and scaling problems that were resolved using EMSC scaling (data not shown).

Exploratory analysis using PCA showed no clear distinction between isolates of different STs (data not shown). However, after further analysis via PC-DFA, it was possible to discriminate between the different subtypes, as samples from the same ST clustered together. The PC-DFA plot from all 95 isolates showed two clear separate clusters (Fig. 1): these two clusters that can easily be distinguished correspond to ST127 and ST131 isolates. The samples from the four remaining STs (ST10, ST69, ST73 and ST95) were not clearly recovered into individual clusters at this stage of the analysis. However, we envisaged that the clustering of samples within the ST131 and ST127 groups may have affected the clustering of the other STs, preventing clear discrimination between the groups; that is to say, the spectra were dominated by specific chemicals characteristic of these two ST groups which were absent from isolates of the other four STs. Therefore, PC-DFA was repeated on isolates of these four STs only, having removed the ST127 (n=9) and ST131 (n=50) data. The resulting plot showed clear clustering of isolates from the remaining STs (Fig. 2).

To validate the observations, the raw spectra were normalised using SNV transformation and subjected to PDS to remove any unavoidable variations between the different ZnSe plates. Prior to correction, the PC-DFA showed clear clustering for ST131 and ST95, and although general trends could be seen for isolates from ST10, ST69, ST73 and ST127, they failed to form distinct clusters (data not shown). After transformation, when the data from plate 2 (test data) were projected into the results from plate 1 (training data), as expected, the ST131 test data projected very close to the

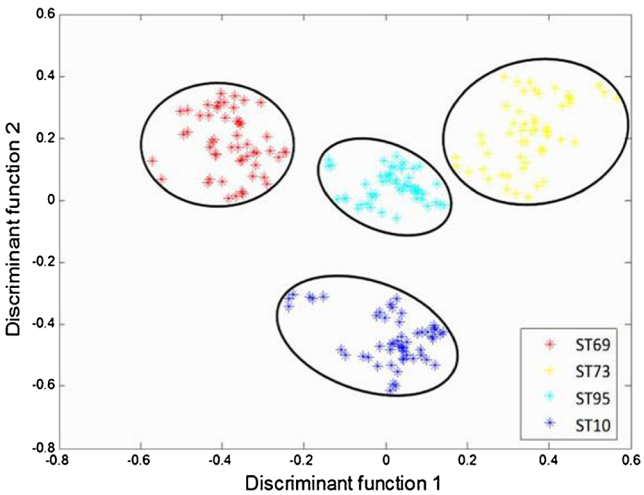

Fig. 2 PC-DFA plot of high-throughput FT-IR spectra for UPEC after the removal of spectra from isolates of ST127 and 131. PCs 1–20 (accounting for 99.54 % of the total variance) were used by the DFA algorithm, again with a priori knowledge of machine replicates (nine isolates from ST10, ST69, ST73 and ST95 were used). Each different colour represents a different ST, as shown by the figure legend. Clear separate clustering for isolates of all four STs can be seen. The *circles* are drawn as a guide and have no statistical significance

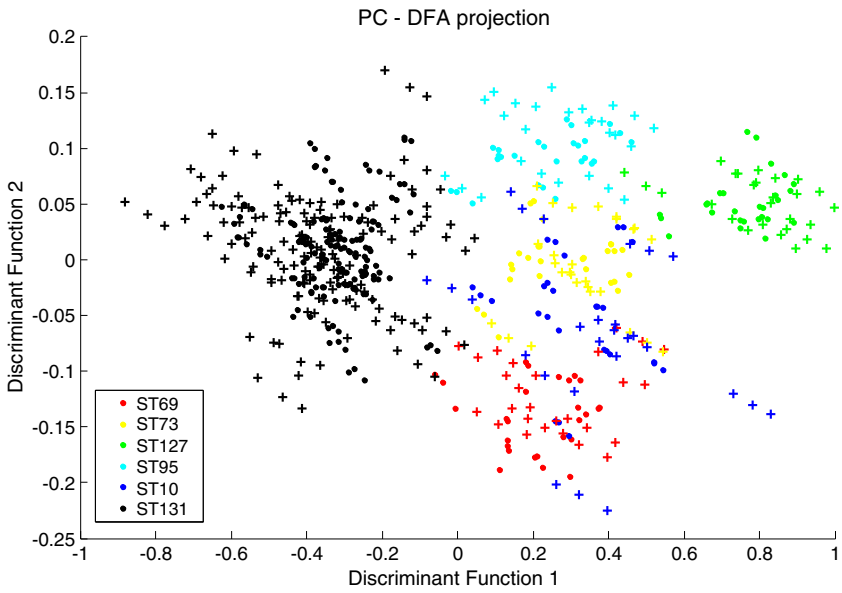

Fig. 3 PC-DFA plot of piecewise direct standardisation (PDS) calibrated data, PC-DFA projection of plate 2 data (test set with the symbol ".") onto plate 1 results (training set with the symbol "x")

training data for this ST (Fig. 3). Similarly, but to a slightly lesser extent, the projections of the ST127, ST73, ST95 and ST69 test data projected close to the training data for the respective STs. Unfortunately, the ST10 projections were not as good as the other STs and many data points appeared to be dispersed away from their respective samples used in the model construction (the training data).

Finally, PLS-DA modelling with 1,000 bootstrapping validations was then performed on the calibrated data and the prediction accuracy for ST131 was 91.19 % (Table 1). Prediction accuracies for ST95 (86.58 %) and ST127 (69.38 %) were also high, showing that it was possible to predict isolates from these groups accurately. Although ST73 (39.15 %) and ST10 (30.15 %) obtained relatively low prediction accuracies, they were better than a truly random prediction, which would have a 1/6th chance (or 16.67 %) of success.

Table 1 Partial least squares-discriminant analysis (PLS-DA) modelling with 1,000 bootstrapping validations showing the prediction accuracies (%) for the isolates examined from six major uropathogenic *Escherichia coli* (UPEC) clones

Actual ST	Predicted ST (%)					
	ST69	ST73	ST127	ST95	ST10	ST131
ST69	55.14	16.65	6.76	5.54	5.58	10.33
ST73	24.71	39.15	6.7	6.99	12.48	9.97
ST127	5.21	8.62	69.38	0.51	7.42	8.85
ST95	1.02	1.67	0.21	86.58	0.15	10.39
ST10	16.48	16.16	17.59	5.58	30.15	14.04
ST131	2.31	0.9	2.41	2.74	0.46	91.19

Discussion

The ability to identify UPEC to the ST level rapidly using FT-IR spectroscopy could help improve the diagnostics and epidemiological surveillance for these common UTI pathogens. A collection of 95 UPEC isolates was examined using HTS FT-IR spectroscopy in order to investigate whether or not it was possible to discriminate between isolates of different STs.

The PCA results obtained from the FT-IR data failed to find six groups, one for each ST, demonstrating the limitation of this unsupervised method in discriminating between these UPEC strains. It was, therefore, necessary to perform PC-DFA to be able to differentiate between these STs. The PC-DFA cluster analysis readily demonstrated the capability of FT-IR to identify UPEC from ST131 (Fig. 1), which agrees with our recent findings using a small collection of UPEC [11]. These current data also show that FT-IR was able to separate ST127 strains from other STs, including ST131. Removing the strong (bio)chemical signal(s) generated by the 50 ST131 strains and the ST127 strains allowed differentiation between ST10, ST69, ST73 and ST95 using PC-DFA and resulted in tight, well-separated clusters (Fig. 2). Being able to identify ST131 and ST127 easily from other STs with FT-IR shows that there are obvious differences in the biochemical fingerprints of these UPEC STs. It is already known that *E. coli* strains can have considerable biochemical variability [24] and we have demonstrated that ST131 and ST127 show significant differences in their biochemical and virulence factor profiles when compared to each other and members of other STs. In combination, this perhaps helps explain the clear separation of isolates from these STs [15, 25].

The between-plate validation carried out resulted in excellent projection for ST131 isolates, with very little variation

between the two biological replicates across the two ZnSe plates. This shows that it should be possible to accurately predict if an isolate is from ST131. However, the projections for ST10 were rather scattered and much less convincing. Whilst the PLS-DA modelling showed very good prediction accuracy for isolates of ST131 (91.19 %), ST95 (86.58 %) and ST127 (69.38 %), the prediction accuracies for ST73 and ST10 were both low (39.15 % and 30.15 %, respectively), which could reflect significant variations between isolates in terms of underlying genetics. It should be noted that there are six different outputs in the PLS-DA model for the prediction of each of the six different bacteria; as such, a truly random score would be a 1 in 6 chance or 16.67 % of being correctly identified. Therefore, as all predictions were much higher than this, one can consider that the bacteria in all of these STs do have unique phenotypic FT-IR fingerprints. eBURST analysis of MLST data has demonstrated that ST10 and ST73 are less recently diverged than other STs, and each is a founder of a large clonal complex, with allelic similarity to many other STs. The deeper divergence may explain why they do not have well-conserved phenotypic characteristics, like members of ST131 and ST127 [15].

The high prediction accuracy for ST131 isolates is particularly promising and supports our previous data for ST131 [11]. ST131 is particularly prevalent worldwide and considered clonal [1, 4, 6]; therefore, the use of HTS FT-IR spectroscopy and the subsequent ability of PC-DFA to discriminate between isolates of different UPEC STs, with minimal sample preparation and relatively little data manipulation, could significantly increase the speed of diagnosis and prove to be an invaluable epidemiological tool.

Acknowledgements TG was supported by a studentship from the Libyan government. HR thanks the Saudi Ministry of Higher Education and King Saud University for funding. RG is grateful to the EU Commonsense (http://www.fp7projectcommonsense.eu/) project (Grant 261809), financed by the European Commission under the Seventh Framework Programme for Research and Technological Development.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Lau SH, Reddy S, Cheesbrough J, Bolton FJ, Willshaw G, Cheasty T, Fox AJ, Upton M (2008) Major uropathogenic *Escherichia coli* strain isolated in the northwest of England identified by multilocus sequence typing. J Clin Microbiol 46(3):1076–1080
2. Kucheria R, Dasgupta P, Sacks SH, Khan MS, Sheerin NS (2005) Urinary tract infections: new insights into a common problem. Postgrad Med J 81(952):83–86. doi:10.1136/pgmj.2004.023036
3. Zhang L, Foxman B, Manning SD, Tallman P, Marrs CF (2000) Molecular epidemiologic approaches to urinary tract infection gene discovery in uropathogenic *Escherichia coli*. Infect Immun 68(4): 2009–2015. doi:10.1128/iai.68.4.2009-2015.2000
4. Cagnacci S, Gualco L, Debbia E, Schito GC, Marchese A (2008) European emergence of ciprofloxacin-resistant *Escherichia coli* clonal groups O25:H4-ST 131 and O15:K52:H1 causing community-acquired uncomplicated cystitis. J Clin Microbiol 46(8):2605–2612. doi:10.1128/JCM.00640-08
5. Peirano G, Pitout JD (2010) Molecular epidemiology of *Escherichia coli* producing CTX-M beta-lactamases: the worldwide emergence of clone ST131 O25:H4. Int J Antimicrob Agents 35(4):316–321. doi: 10.1016/j.ijantimicag.2009.11.003
6. Rogers BA, Sidjabat HE, Paterson DL (2011) *Escherichia coli* O25b-ST131: a pandemic, multiresistant, community-associated strain. J Antimicrob Chemother 66(1):1–14. doi:10.1093/jac/dkq415
7. Johnson JR, Menard M, Johnston B, Kuskowski MA, Nichol K, Zhanel GG (2009) Epidemic clonal groups of *Escherichia coli* as a cause of antimicrobial-resistant urinary tract infections in Canada, 2002 to 2004. Antimicrob Agents Chemother 53(7):2733–2739. doi: 10.1128/AAC.00297-09
8. Köhling HL, Bittner A, Müller K-D, Buer J, Becker M, Rübben H, Rettenmeier AW, Mosel F (2012) Direct identification of bacteria in urine samples by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and relevance of defensins as interfering factors. J Med Microbiol 61(Pt 3):339–344. doi:10.1099/jmm.0.032284-0
9. Jarvis RM, Goodacre R (2004) Ultra-violet resonance Raman spectroscopy for the rapid discrimination of urinary tract infection bacteria. FEMS Microbiol Lett 232(2):127–132. doi:10.1016/s0378-1097(04)00040-0
10. Goodacre R, Timmins EM, Burton R, Kaderbhai N, Woodward AM, Kell DB, Rooney PJ (1998) Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks. Microbiology 144:1157–1170
11. AlRabiah H, Correa E, Upton M, Goodacre R (2013) High-throughput phenotyping of uropathogenic *E. coli* isolates with Fourier transform infrared spectroscopy. Analyst 138(5):1363–1369. doi:10.1039/c3an36517d
12. Gilbert MK, Frick C, Wodowski A, Vogt F (2009) Spectroscopic imaging for detection and discrimination of different *E. coli* strains. Appl Spectrosc 63(1):6–13
13. Goodacre R, Timmins EM, Rooney PJ, Rowland JJ, Kell DB (1996) Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance–absorbance Fourier transform infrared spectroscopy and artificial neural networks. FEMS Microbiol Lett 140(2–3): 233–239. doi:10.1111/j.1574-6968.1996.tb08342.x
14. Winder CL, Gordon SV, Dale J, Hewinson RG, Goodacre R (2006) Metabolic fingerprints of *Mycobacterium bovis* cluster with molecular type: implications for genotype–phenotype links. Microbiology 152(Pt 9):2757–2765. doi:10.1099/mic.0.28986-0
15. Gibreel TM, Dodgson AR, Cheesbrough J, Fox AJ, Bolton FJ, Upton M (2012) Population structure, virulence potential and antibiotic susceptibility of uropathogenic *Escherichia coli* from Northwest England. J Antimicrob Chemother 67(2):346–356. doi:10.1093/jac/dkr451
16. Goodacre R (2003) Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. Vib Spectrosc 32(1): 33–45. doi:10.1016/s0924-2031(03)00045-6
17. Nicolaou N, Goodacre R (2008) Rapid and quantitative detection of the microbial spoilage in milk using Fourier transform infrared spectroscopy and chemometrics. Analyst 133(10):1424–1431
18. Barnes RJ, Dhanoa MS, Lister SJ (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl Spectrosc 43:772–777
19. Wang Y, Kowalski BR (1992) Calibration transfer and measurement stability of near-infrared spectrometers. Appl Spectrosc 46:764–771
20. Jolliffe IT (1986) Principal component analysis. Springer-Verlag, New York

21. Jarvis RM, Goodacre R (2004) Discrimination of bacteria using surface-enhanced Raman spectroscopy. Anal Chem 76(1):40–47. doi:10.1021/ac034689c
22. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemometr 17(3):166–173. doi:10.1002/cem.785
23. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26
24. Godbout-DeLasalle F, Higgins R (1986) Biotyping of clinical isolates of *Escherichia coli* of animal origin, using the Analytab API 20E system. Can J Vet Res 50(3):418–421
25. Gibreel TM, Dodgson AR, Cheesbrough J, Bolton FJ, Fox AJ, Upton M (2012) High metabolic potential may contribute to the success of ST131 uropathogenic *Escherichia coli*. J Clin Microbiol 50(10): 3202–3207. doi:10.1128/jcm.01423-12